EXPOSING DEEP FAKES USING INCONSISTENT HEAD POSES

Xin Yang[⋆], *Yuezun Li*[⋆] *and Siwei Lyu*

University at Albany, State University of New York, USA

ABSTRACT

In this paper, we propose a new method to expose AI-generated fake face images or videos (commonly known as the *Deep Fakes*). Our method is based on the observations that Deep Fakes are created by splicing synthesized face region into the original image, and in doing so, introducing errors that can be revealed when 3D head poses are estimated from the face images. We perform experiments to demonstrate this phenomenon and further develop a classification method based on this cue. Using features based on this cue, an SVM classifier is evaluated using a set of real face images and Deep Fakes.

***Index Terms*—** Media Forensics, DeepFake Detection, Head Pose Estimation

1. INTRODUCTION

Thanks to the recent developments of machine learning, the technologies for manipulating and fabricating images and videos have reached a new level of sophistication [1, 2, 3, 4, 5, 6, 7]. The cutting edge of this trend are the so-called Deep Fakes, which are created by inserting faces synthesized using deep neural networks into original images/videos. Together with other forms of misinformation shared through the digital social network, Deep Fakes created digital impersonations have become a serious problem with negative social impacts [8]. Accordingly, there is an urgent need for effective methods to expose Deep Fakes.

To date, detection methods of Deep Fakes have relied on artifacts or inconsistencies intrinsic to the synthesis algorithms, for instance, the lack of realistic eye blinking [9], mismatched color profiles [10] and visual lips with speeches [11]. Neural network based classification approach has also been used to directly discern real imagery from Deep Fakes [12, 13, 14]. In this work, we propose a new approach to detect Deep Fakes. Our method is based on an intrinsic limitation in the deep neural network face synthesis models, which is the core component of the Deep Fake production pipeline. Specifically, these algorithms create faces of a different person but keeping the facial expression of the original person. However, the two faces have mismatched facial landmarks, which are locations on human faces corresponding to important structures such as eye and mouth tips, as the neural network synthesis algorithm does not guarantee the original face and the synthesized face to have consistent facial landmarks, as shown in Fig. 1.

The errors in landmark locations may not be visible directly to human eyes, but can be revealed from head poses (*i.e*, head orientation and position) estimated from 2D landmarks in the real and faked parts of the face. Specifically, we compare head poses estimated using all facial landmarks and those estimated using only the central region, as shown in Fig. 1. The rationale is that the two estimated head poses will be close for the original face, Fig. 1(k). But for a Deep Fake, as the central face region is from the synthesized face, the errors due to the mismatch of landmark locations from original and generated images aforementioned will lead to a larger difference between the two estimated head poses, Fig. 1(n). We experimentally confirm the significant difference in the estimated head poses in Deep Fakes. Then we use the difference in estimated head poses as a feature vector to train a simple SVM based classifier to differentiate original and Deep Fakes. Experiments on realistic Deep Fake videos demonstrate the effectiveness of our algorithm.

2. DEEP FAKE PRODUCTION PIPELINE

The overall process of making Deep Fakes is illustrated in Fig. 1(a) - (h). To generate a Deep Fake, we feed the algorithm an image (or a frame from a video) that contains the source face to be replaced. A bounding box of this face is obtained with a face detector, followed by the detection of facial landmarks. The face area is warped into a standard configuration through an affine transformation M, by minimizing the alignment errors of central facial landmarks (red dots in Fig. 1(c)) to a set of standard landmark locations, a process known as face alignment. This image is then cropped into 64×64 pixels, and fed into the deep generative neural network to create a synthesized face. The synthesized face is transformed back with M^{-1} to match the original face. Finally, with post-processing such as boundary smoothing, a Deep Fake image/video frame is created.

[⋆] The authors contribute equally.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

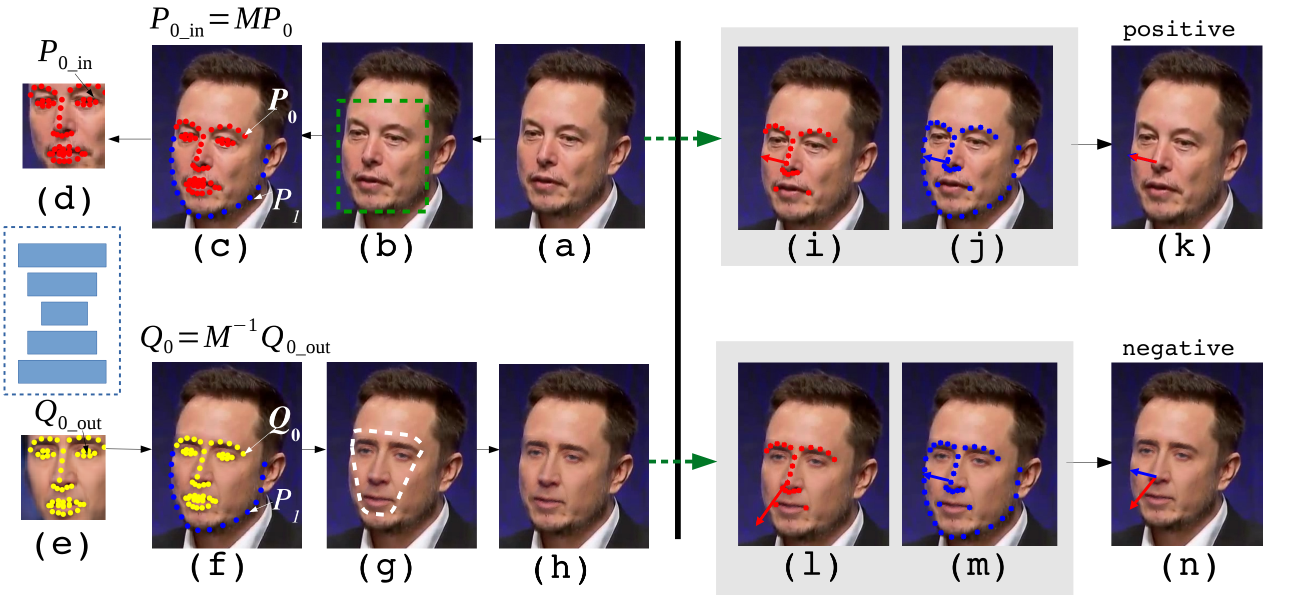

Fig. 1. *Overview of Deep Fake work-flow (Left) and our method (Right). In* **(Deep Fake work-flow)***:* **(a)** *is the original image.* **(b)** *Detected face in the image.* **(c)** *Detected 2D facial landmarks.* **(d)** *Cropped face in (a) is warped to a standardized face using an affine transformation* M*.* **(e)** *Deep Fake face synthesized by the deep neural network.* **(f)** *Deep Fake face is transformed back using* M^{-1}*.* **(g)** *The mask of transformed face is refined based on landmarks.* **(g)** *Synthesized face is merged into the original image.* **(h)** *The final fake image. For* **(our method)***: The top row corresponds to a real image and the bottom corresponds to a Deep Fake. We compare head poses estimated using facial landmarks from the whole face* **(j)**, **(m)** *or only the central face region* **(i)**, **(l)**. *The alignment error is revealed as differences in the head poses shown as their projections on the image plane. The difference of the head poses is then fed to an SVM classifier to differentiate the original image* **(k)** *from the Deep Fake* **(n)**.

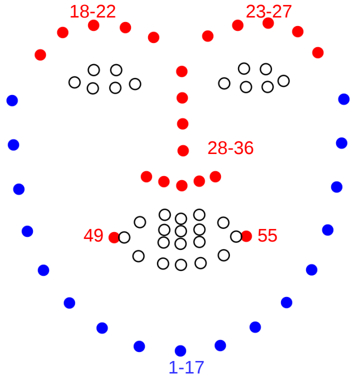

Fig. 2. 68 *facial landmarks. Red dots are used as central face region. Blue and red landmarks are used as whole face. The landmarks represented as empty circles are not used in head pose estimation.*

3. 3D HEAD POSE ESTIMATION

The 3D head pose corresponds to the rotation and translation of the world coordinates to the corresponding camera coordinates. Specifically, denote $[U, V, W]^T$ as the world coordinates of one facial landmark, $[X, Y, Z]^T$ be its camera coordinates, and $(x, y)^T$ be its image coordinates. The transformation between the world and the camera coordinate systems can be formulated as

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = R \begin{bmatrix} U \\ V \\ W \end{bmatrix} + \vec{t}, \tag{1}$$

where R is the 3×3 rotation matrix, $\vec{t}$ is 3×1 translation vector. The transformation between camera and image coordinate systems is defined as

$$s \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{2}$$

where f_x and f_y are the focal lengths in the x- and y-directions and (c_x, c_y) is the optical center, and s is an unknown scaling factor.

In 3D head pose estimation, we need to solve the reverse problem, *i.e*, estimating s, R and $\vec{t}$ using the 2D image coordinates and 3D world coordinates of the same set of facial landmarks obtained from a standard model, *e.g*, a 3D average face model, assuming we know the camera parameter. Specifically, for a set of n facial landmark points, this can be formulated as an optimization problem, as

$$\min_{R,\vec{t},s} \sum_{i=1}^{n} \left\| s \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} - \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \left(R \begin{bmatrix} U_i \\ V_i \\ W_i \end{bmatrix} + \vec{t} \right) \right\|^2$$

that can be solved efficiently using the Levenberg-Marquardt algorithm [15]. The estimated R is the camera pose which is the rotation of the camera with regards to the world coordinate, and the head pose is obtained by reversing it as R^T (as R is an orthornormal matrix).

4. INCONSISTENT HEAD POSES IN DEEP FAKES

As a result of swapping faces in the central face region in the Deep Fake process in Fig. 1, the landmark locations of fake faces often deviate from those of the original faces. As shown in Fig. 1(c), a landmark in the central face region P_0 is firstly affine-transformed into $P_{0\_in} = MP_0$. After the generative neural network, its corresponding landmark on the faked face is Q_{0_out}.

As the configuration of the generative neural network in Deep Fake does not guarantee landmark matching, and people have different facial structures, this landmark Q_{0_out} on generated face could have different locations to P_{0_in}. Based on the comparing 51 central region landmarks of 795 pairs of images in 64 × 64 pixels, the mean shifting of a landmark from the input (Fig. 1(d)) to the output (Fig. 1(e)) of the generative neural network is 1.540 pixels, and its standard deviation is 0.921 pixel. After an inverse transformation $Q_0 = M^{-1}Q_{0\_out}$, the landmark locations Q_0 in the faked faces will differ from the corresponding landmarks P_0 in the original face. However, due to the fact that Deep Fake only swap faces in the central face region, the locations of the landmarks on the outer contour of the face (blue landmarks P_1 in Fig. 1(c) and (f)) will remain the same. This mismatch between the landmarks at center and outer contour of faked faces is revealed as inconsistent 3D head poses estimated from central and whole facial landmarks. Particularly, the head pose difference between central and whole face region will be small in real images, but large in fake images.

We conduct experiments to confirm our hypothesis. For simplicity, we look at the head orientation vector only. Denote R_a^T as the rotation matrix estimated using facial landmarks from the whole face (red and blue landmarks in Fig. 2) using method described in Section 3, and R_c^T as the one estimated using only landmarks in the central region (red landmarks in Fig. 2). We obtain the 3D unit vectors $\vec{v}_a$ and $\vec{v}_c$ corresponding to the orientations of the head estimated this way, as $\vec{v}_a = R_a^T\vec{w}$ and $\vec{v}_c = R_c^T\vec{w}$, respectively, with $\vec{w} = [0, 0, 1]^T$ being the direction of the w-axis in the world coordinate. We then compare the cosine distance between the two unit vectors $\vec{v}_c$ and $\vec{v}_a$, $1 - \vec{v}_a \cdot \vec{v}_c/(\|\vec{v}_a\|\|\vec{v}_c\|)$, which takes value in $[0, 2]$ with 0 meaning the two vectors agree with each other. The smaller this value is, the closer the two vectors are to each other. Shown in Fig. 3 are histograms of the cosine distances between $\vec{v}_c$ and $\vec{v}_a$ for a set of original and Deep Fake generated images. As these results show, the cosine distances of the two estimated head pose vectors for the real images concentrates on a significantly smaller range of values up to 0.02, while for Deep Fakes the majority of the values are in the range between 0.02 and 0.08. The difference in the distribution of the cosine distances of the two head orientation vectors for real and Deep Fakes suggest that they can be differentiated based on this cue.

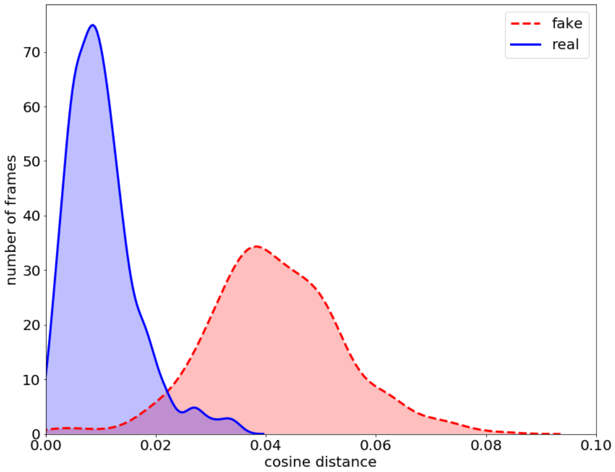

Fig. 3. *Distribution of the cosine distance between $\vec{v}_c$ and $\vec{v}_a$ for fake and real face images.*

5. CLASSIFICATION BASED ON HEAD POSES

We further trained SVM classifiers based on the differences between head poses estimated using the full set of facial landmarks and those in the central face regions to differentiate Deep Fakes from real images or videos. The features are extracted in following procedures: (1) For each image or video frame, we run a face detector and extract 68 facial landmarks using software package `DLib` [16]. (2) Then, with the standard 3D facial landmark model of the same 68 points from OpenFace2 [17], the head poses from central face region (R_c and t_c) and whole face (R_a and t_a) are estimated with landmarks $18 - 36, 49, 55$ (red in Fig. 2) and $1 - 36, 49, 55$ (red and blue in Fig. 2), respectively. Here, we approximate the camera focal length as the image width, camera center as image center, and ignore the effect of lens distortion. (3) The differences between the obtained rotation matrices $(R_a - R_c)$ and translation vectors $(\vec{t}_a - \vec{t}_c)$ are flattened into a vector, which is standardized by subtracting its mean and divided by its standard deviation for classification.

The training and testing data for the SVM classifier are based on two datasets of real and Deep Fake images and videos. The first, `UADFV`, is a set of Deep Fake videos and their corresponding real videos that are used in our previous work [9]. This dataset contains 49 real videos, which were used to create 49 Deep Fake videos. The average length of these videos is approximately 11.14 seconds, with a typical resolution of 294 × 500 pixels. The second dataset is a subset from the DARPA MediFor GAN Image/Video Challenge [18] , which has 241 real images and 252 Deep Fake images. For the training of the SVM classifier, we use frames from 35 real and 35 Deep Fake videos in the UADFV dataset, with a total number of $21,694$ images. Frames (a total $11,058$ frames) from the remaining 14 real and 14 Deep Fake videos from the UADFV dataset and all images in the DARPA GAN set are used to test the SVM classifiers. We train SVM classifier with RBF kernels on the training data, with a grid search on the hyperparameters using 5-fold cross validation.

The performance, evaluated using individual frames as unit of analysis with Area Under ROC (AUROC) as the performance metric, is shown for the two datasets in Fig. 4. As these results show, on the UADFV dataset, the SVM classifier achieves an AUROC of 0.89. This indicates that the difference between head poses estimated from central region and whole face is a good feature to identify Deep Fake generated images. Similarly, on the DARPA GAN Challenge dataset, the AUROC of the SVM classifier is 0.843. This results from the fact that the synthesized faces in the DARPA GAN challenges are often blurry, leading to difficulties to accurately predict facial landmark locations, and consequently the head pose estimations. We also estimate the performance using individual videos as unit of analysis for the UADFV dataset. This is achieved by averaging the classification prediction on frames over individual videos. The performance is shown in the last row of Table 1.

We also perform an ablation study to compare the performance of different types of features used in the SVM classifier. Specifically, we compare $five$ different types of features based on the rotation and translation of estimated 3D head poses in camera coordinates are also examined as in Table 1. (1) As in Section 4, we simplified head poses as head orientations, $\vec{v}_a$ and $\vec{v}_c$. Classification using $\vec{v}_a - \vec{v}_c$ as features achieves 0.738 AUROC on Deep Fake Dataset. This is expected, as this simplification neglects the translation and rotation on other axes. (2) As there are 3 degrees of freedom in rotation, representing head pose rotation matrix as Rodrigues' rotation vector ($\vec{r}_a - \vec{r}_c$) could increase the AUROC to 0.798. (3) Instead of Rodrigues' vector $\vec{r} \in R^3$, flatten the difference of 3 by 3 rotation matrices $R_a - R_c$ as features further improves the AUROC to 0.840. (4) Introducing the difference of translation vectors $\vec{t}_a - \vec{t}_c$ to (1) and (2) results in AUROCs as 0.866 and 0.890, due to the increase of head poses in translation.

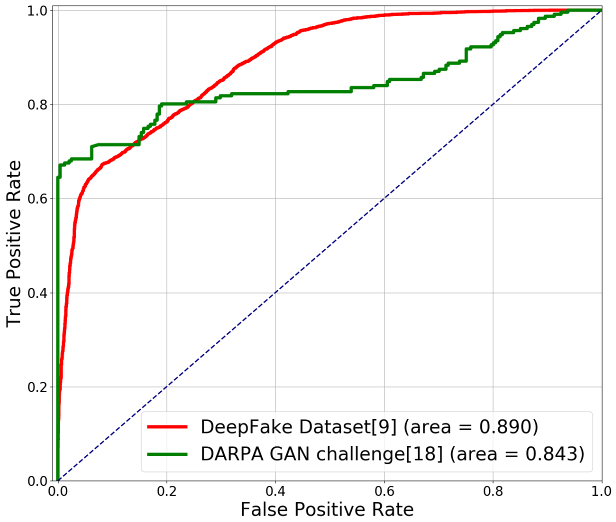

Fig. 4. *ROC curves of the SVM classification results, see texts for details.*

Table 1. AUROC based on videos and frames from UADFV dataset [9]

features	frame	video
$\vec{v}_a - \vec{v}_c$	0.738	0.888
$\vec{r}_a - \vec{r}_c$	0.798	0.898
$R_a - R_c$	0.853	0.913
$(\vec{v}_a - \vec{v}_c)$ & $(\vec{t}_a - \vec{t}_c)$	0.840	0.949
$(\vec{r}_a - \vec{r}_c)$ & $(\vec{t}_a - \vec{t}_c)$	0.866	0.954
$(R_a - R_c)$ & $(\vec{t}_a - \vec{t}_c)$	0.890	0.974

6. CONCLUSION

In this paper, we propose a new method to expose AI-generated fake face images or videos (commonly known as the *Deep Fakes*). Our method is based on observations that such Deep Fakes are created by splicing a synthesized face region into the original image, and in doing so, introducing errors that can be revealed when 3D head poses are estimated from the face images. We perform experiments to demonstrate this phenomenon and further develop a classification method based on this cue. We also report experimental evaluations of our methods on a set of real face images and Deep Fakes.

7. REFERENCES

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arxiv*, 2016.

[2] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3677–3685.

[3] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni, "On face segmentation, face swapping, and face perception," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 98–105.

[4] Hai X Pham, Yuting Wang, and Vladimir Pavlovic, "Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network," *arXiv preprint arXiv:1803.07716*, 2018.

[5] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, N. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep Video Portraits," *ACM Transactions on Graphics 2018 (TOG)*, 2018.

[6] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2387–2395.

[7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[8] Robert Chesney and Danielle Keats Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *107 California Law Review (2019, Forthcoming); U of Texas Law, Public Law Research Paper No. 692; U of Maryland Legal Studies Research Paper No. 2018-21*.

[9] Yuezun Li, Ming-Ching Chang, and Siwei Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[10] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang, "Detection of deep network generated images using disparities in color components," *arXiv preprint arXiv:1808.07276*, 2018.

[11] Pavel Korshunov and Sébastien Marcel, "Speaker inconsistency detection in tampered video," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2375–2379.

[12] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Mesonet: a compact facial video forgery detection network," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[13] Pavel Korshunov and Sébastien Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.

[14] David Güera and Edward J Delp, "Deepfake video detection using recurrent neural networks," in *IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS)*, 2018.

[15] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[16] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[17] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 59–66.

[18] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N. Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus, "MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation," in *IEEE Winter Conf. on Applications of Computer Vision (WACV), Workshop on Image and Video Forensics*, 2019.